Classification Algorithms for SIFT-MS Medical Diagnosis

K. Moorhead, D. Lee, J.G. Chase, A. Moot, K. Ledingham, J. Scotter, R. Allardyce, S. Senthilmohan, Z. Endre

Katherine Moorhead

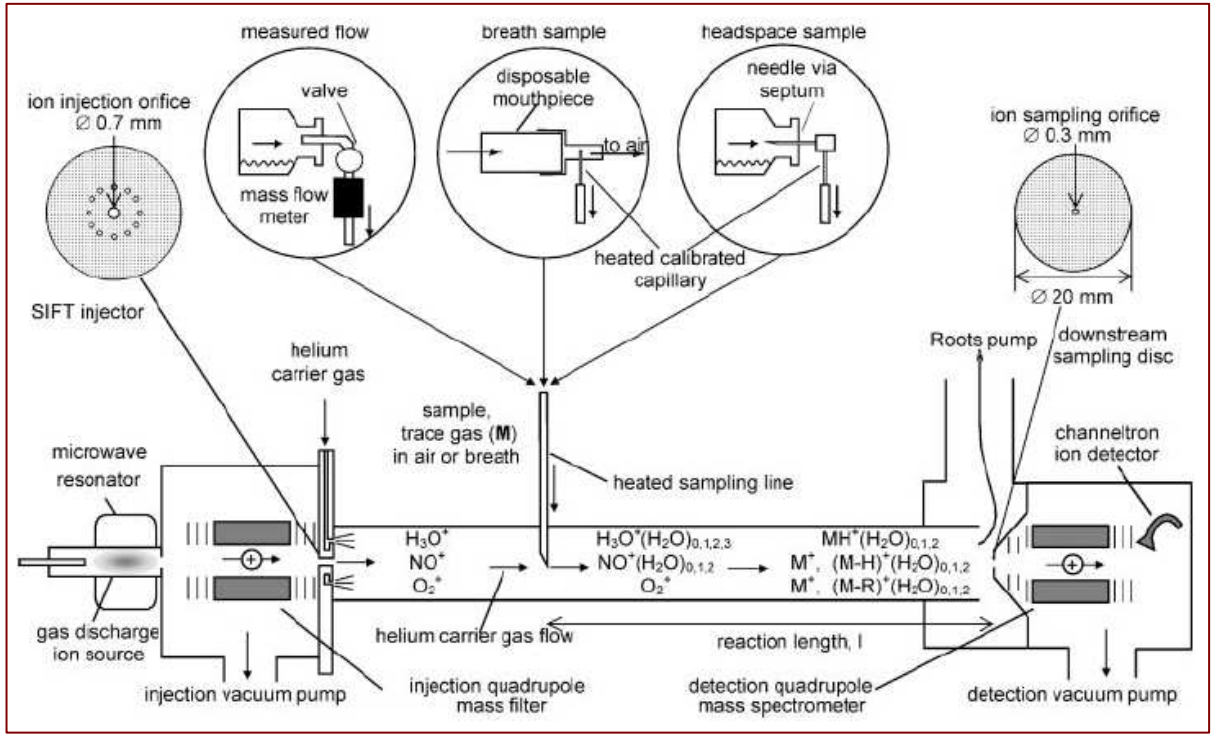
University of Canterbury

Christchurch

New Zealand

SIFT-MS

- Selected Ion Flow Tube Mass Spectrometry
- Chemical ionisation of positively charged precursor ions that react with the VOCs in air/breath sample



1. Precursor ions generated
2. Quadrupole mass filter selects required precursor
3. Precursor ion injected into inert carrier gas
4. Reaction of VOCs and precursor
5. Product ions filtered according to mass
6. Product ions counted in channeltron particle multiplier/detector

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Introduction

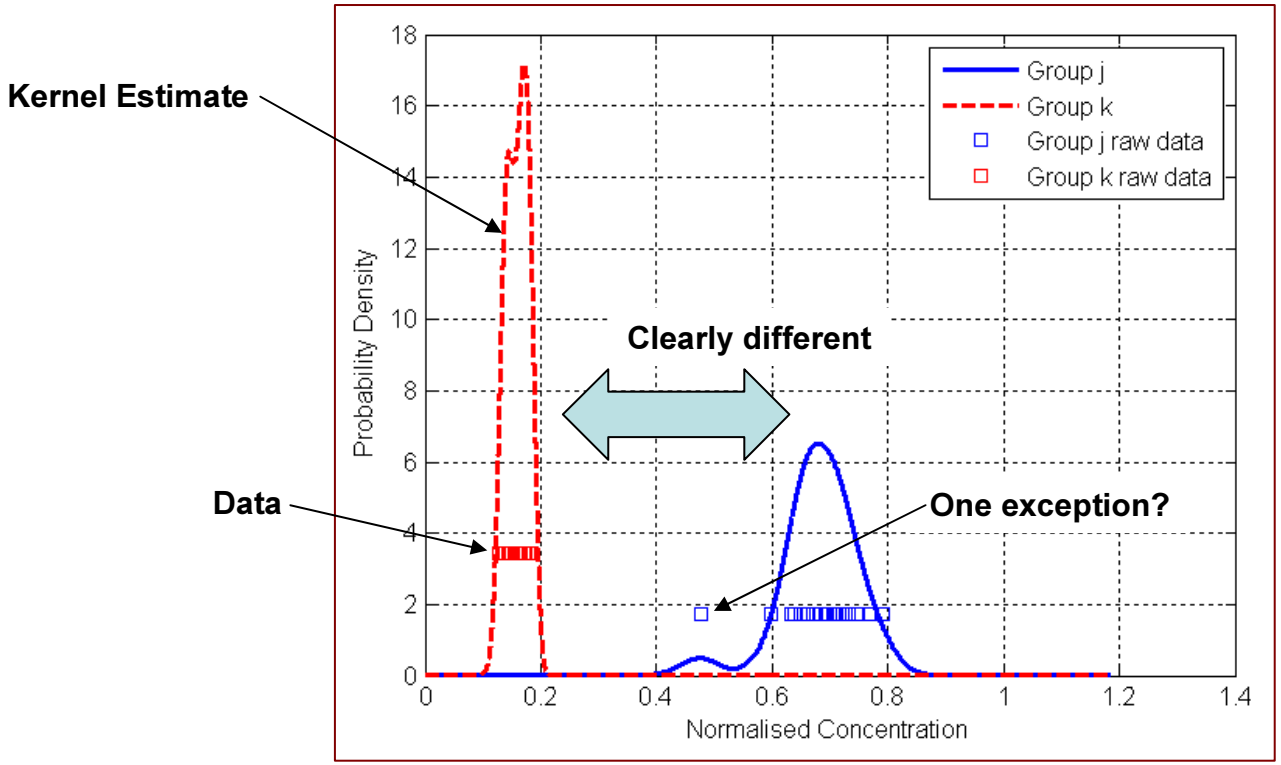
- SIFT-MS offers real-time quantification of trace gases
- The SIFT-MS system can potentially offer unique clinical capabilities
  - –Early and rapid detection of disease, infectious bacteria and/or patient condition would have significant clinical impact
- **Requires**: a classifier to differentiate between control and test groups.
- **Goal**: classification and/or identification of the masses and VOCs that contribute most strongly towards a successful classification
- **Outcome**: may allow new or improved biomarkers for a particular disease state or patient condition to be discovered

Classification

- Develop database and classify unknown sample into a database group
- Identify masses that contribute towards a successful classification and thus act as good biomarkers
- Use paired data to determine differences over time
- **Overall there are two approaches**:
  - **Classification + Biomarkers** = a “pure” classification problem assuming independent patients and data
  - **Biomarker ID only** = uses paired or grouped patient data to ID changes over state or time

Kernel Density Estimate

- Mixed distribution made up of a kernel density and a Dirac delta function is used to develop a density profile at **each mass**



UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UCBIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

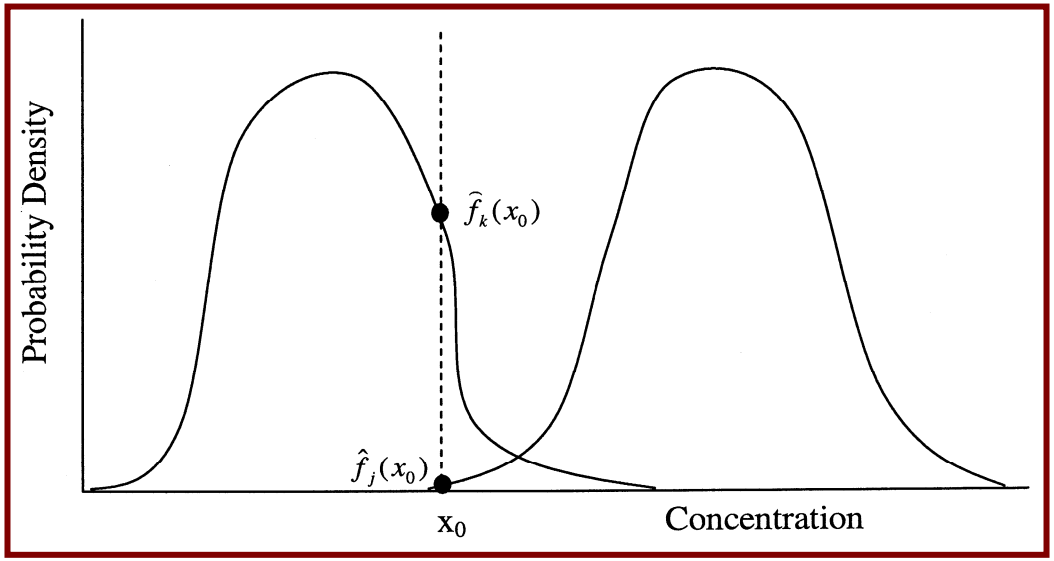
syft™ Technologies

Classifying unknown sample

- Probability of sample being in Group *j* given x_0

$$\hat{\Pr}(j \mid x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\hat{\pi}_k \hat{f}_k(x_0) + \hat{\pi}_j \hat{f}_j(x_0)}$$

- If ratio is greater than *q*, sample is classified in Group *j*



Log-odds Ratio

- Log-odds ratio obtained for each mass

$$LogoddsRatio = \ln\left(\frac{\hat{\Pr}(j \mid x_0)}{\hat{\Pr}(k \mid x_0)}\right)$$

- Log-odds ratio 'summed' over all masses

$$\ln\frac{\hat{\Pr}(j \mid x_0)}{\hat{\Pr}(k \mid x_0)} = \ln\frac{\hat{\pi}_j}{\hat{\pi}_k} + \sum_{m=1}^{M} \ln\frac{\hat{f}_j(x_0., m)}{\hat{f}_k(x_0., m)}$$

- Ratio greater than ln[q/(1-q)] indicates the sample is in the numerator group (Group *j)*
- *[If q = 0.5 and probability of Group k classification more likely at most masses, then the term:* $\ln\left(\frac{\hat{\Pr}(j \mid x_0)}{\hat{\Pr}(k \mid x_0)}\right)$ *is negative]*

Prediction Error Estimation

- Stratified bootstrap method
  - Choose with replacement from original sample
  - Bootstrap sample created of same size as the original sample
  - Repeated *B* times (~1000)

$$\hat{E}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} I(\hat{y}^b(x_i) \neq y_i)$$

- Alleviate bias using 0.632 estimator

$$\hat{E}^{(0.632)} = 0.368 \times \bar{e} + 0.632 \times \hat{E}^{(1)}$$

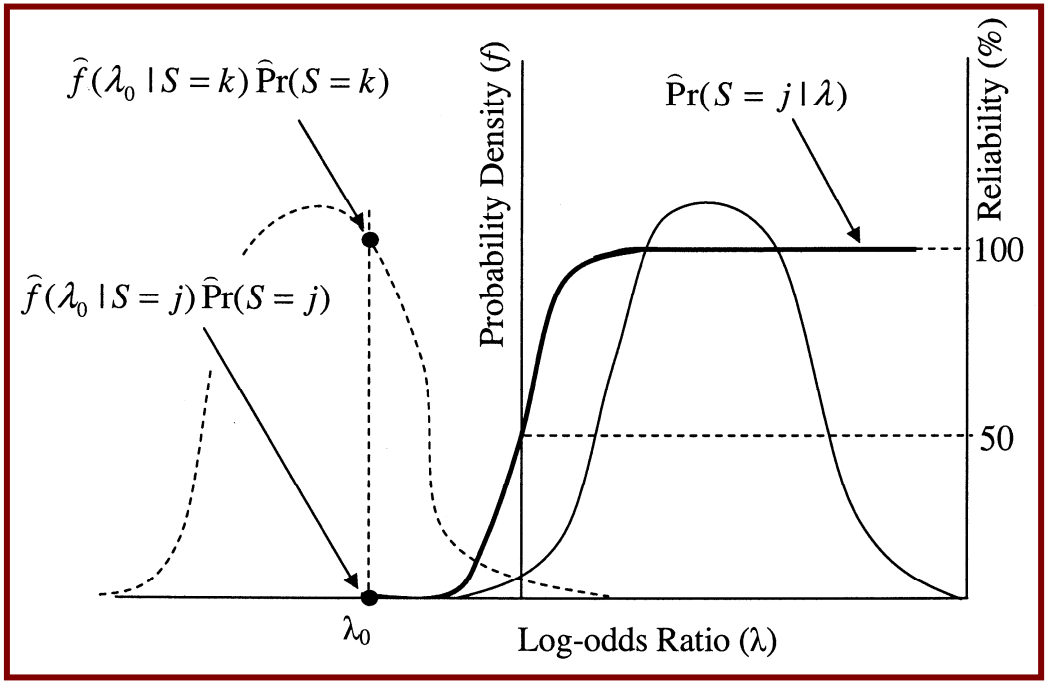
- $\bar{e}$ is the biased error using all data as training set

- *N* is the total number of samples
- $|C^{-i}|$ is the number of bootstrap samples that do not contain sample *i*
- $\hat{y}^b$ is the classifier trained on bootstrap sample *b*
- $I(\hat{y}^b(x_i) \neq y_i)$ equals 1 if classified incorrectly
- y_i is the group x_i belongs to

Reliability

- Create density profiles of the log-odds ratios obtained

$$\hat{\Pr}(S = j \mid \lambda) = \frac{\hat{f}(\lambda \mid S = j)\,\hat{\Pr}(S = j)}{\hat{f}(\lambda \mid S = j)\,\hat{\Pr}(S = j) + \hat{f}(\lambda \mid S = k)\,\hat{\Pr}(S = k)}$$

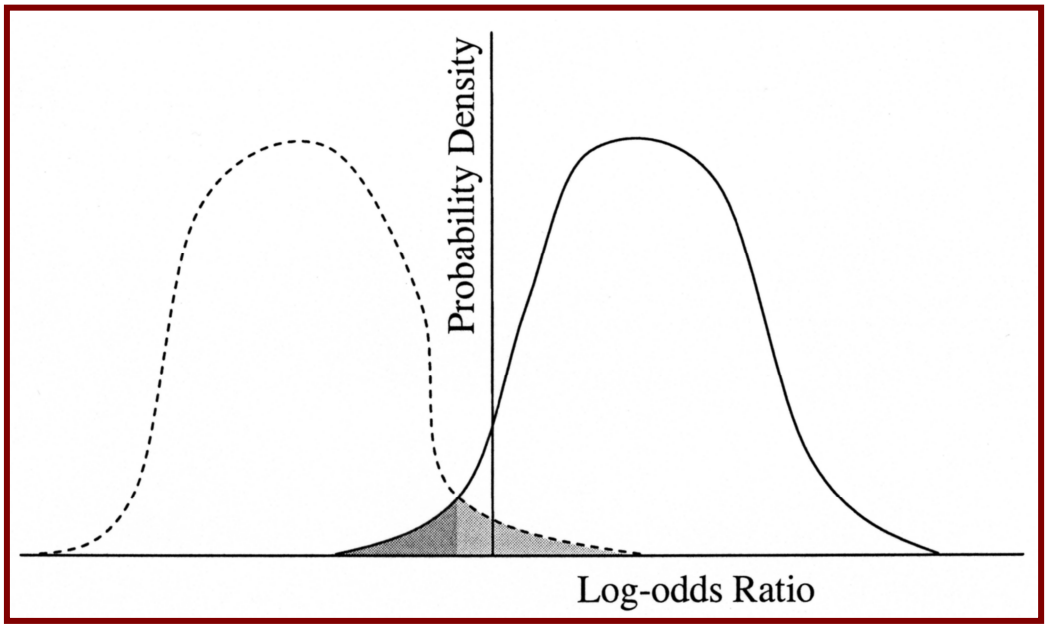


ROC Curve

- The null hypothesis is that the unknown sample is in Group *j*.
- Non-specificity = the probability of incorrectly rejecting the null hypothesis (classifying the sample as Group *k* when in fact it is Group *j*)
- Sensitivity = the probability of correctly rejecting the null hypothesis (classifying the sample as Group *k* when in fact it is Group *k*).

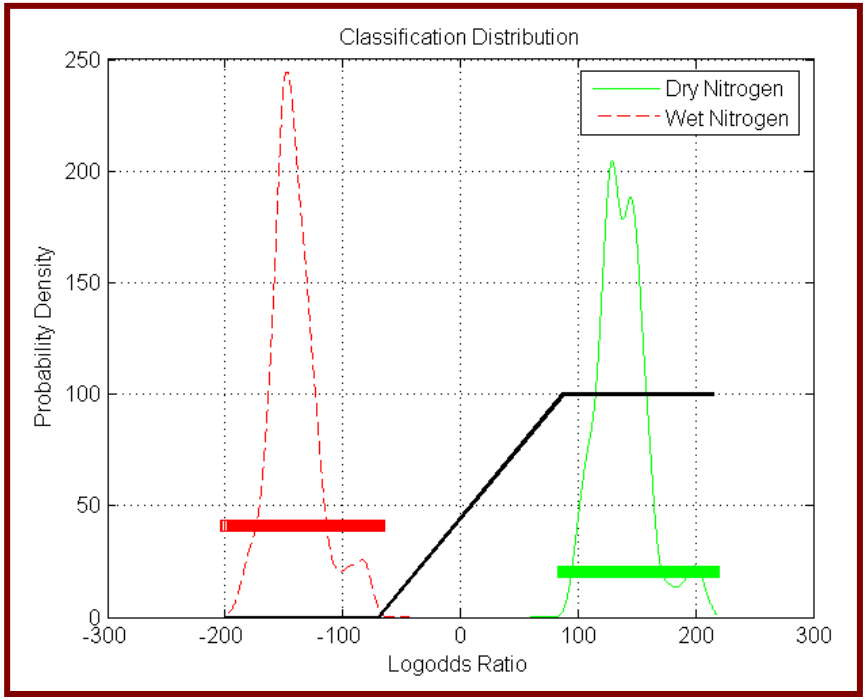
Biomarkers – A simple method

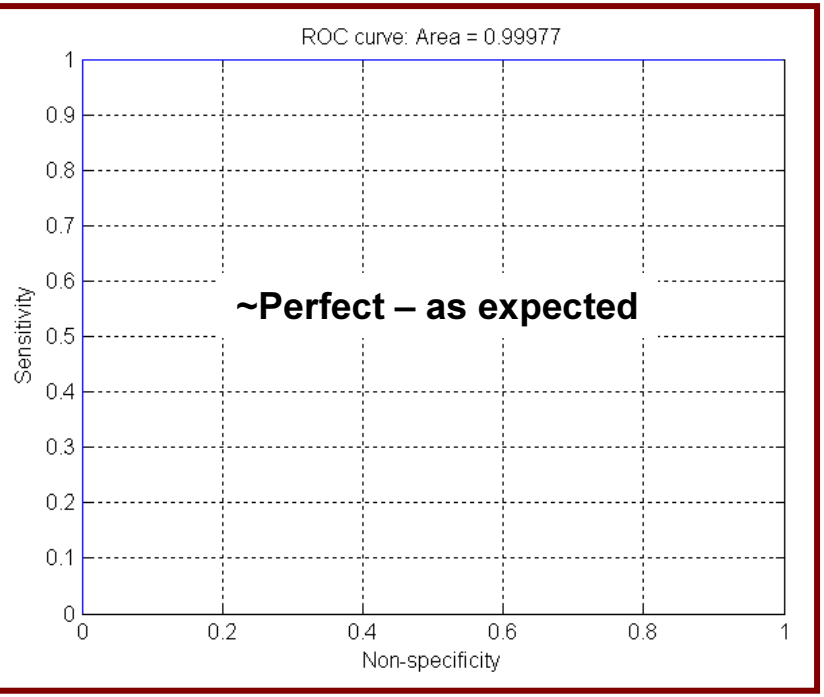
- **Biomarker Query** –which masses contain minimal overlap (maximum separation) between log-odds density profiles?



Tedlar Bag “Idiot” Validation

- 25 nitrogen samples collected in a tedlar bag
- 25 nitrogen samples vented to sterile glass bottles
- Probability density profiles use normalised concentration values





UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UCBIOENGINEERING

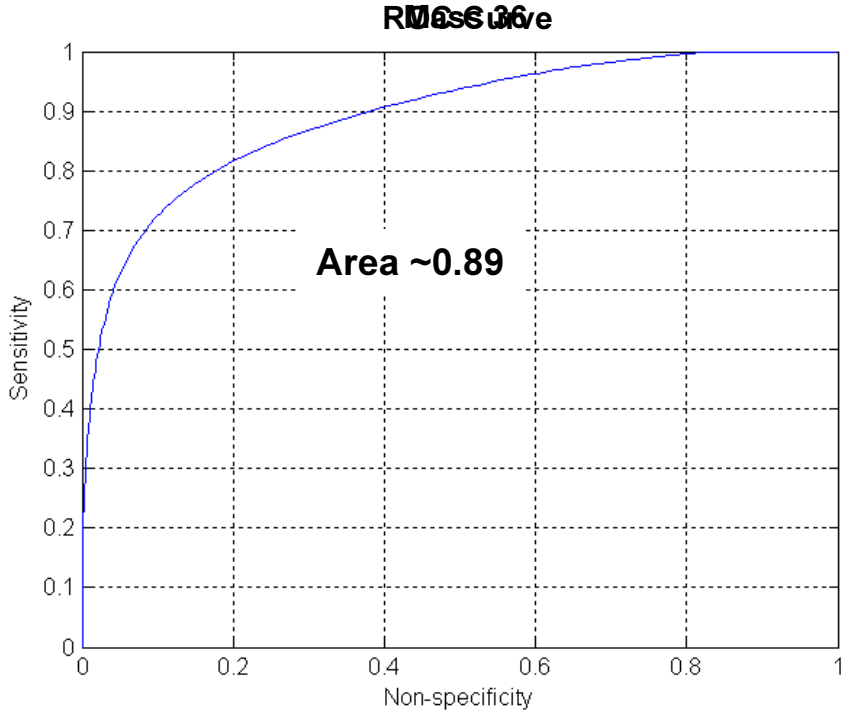
Canterbury District Health Board Te Poari Hauora ō Waitaha

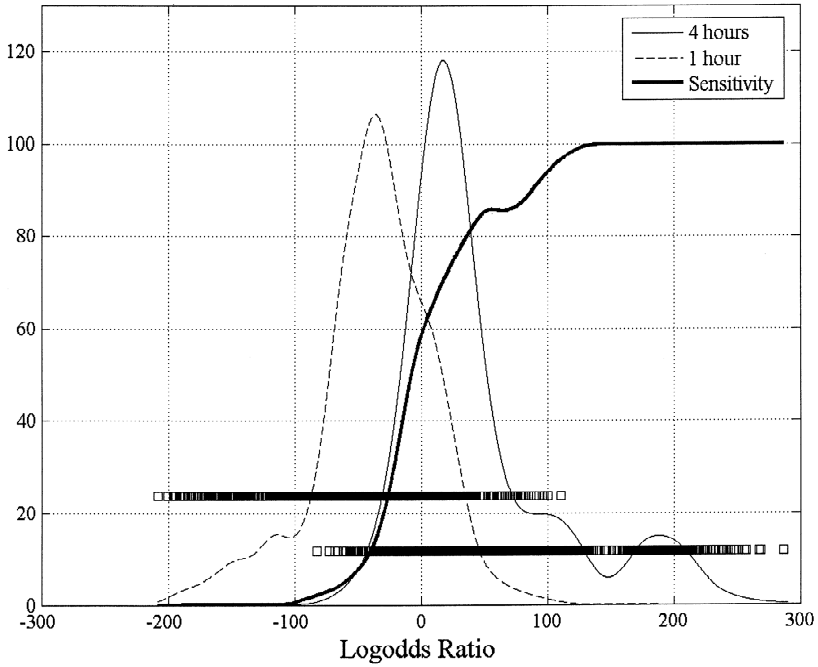
Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Dialysis Study Proof of Concept

- 7 repeat mass scans were taken of 1 dialysis patient
- Mass scans taken at 1 and 4 hours = 14 total points (7 each group)





UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Paired Data Approach

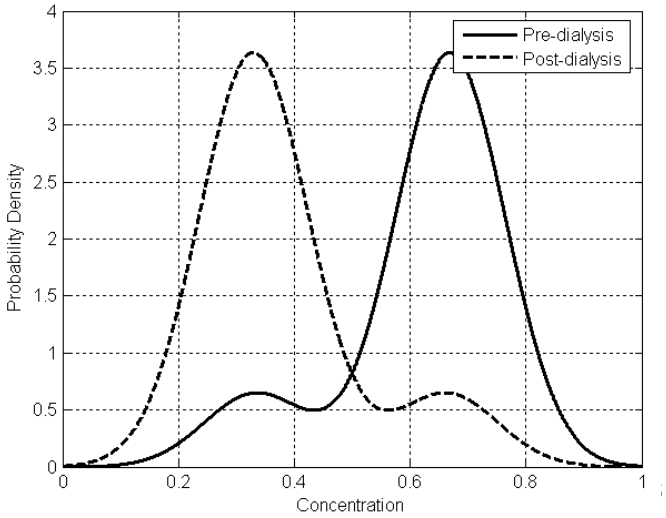
- **Previous method**: focus on differentiating between 2 distinct groups
- **Goal**: discover which masses, *m*, change the most between 2 chosen time points, *j* and *k*, and thus act as good biomarkers for kidney function.
- Why not examine variations in patients before and after treatment?
  - Paired data where data point in the pre-dialysis group (Group *k*) is linked to a specific data point in the post-dialysis group (Group *j*).
- What trade-off can we find between adequate sample sizes and patient specific responses?
  - Patients too variable to lump together (different temporal profile, starting and equilibrium concentration)
  - Each patients dataset is too small to be considered independently
- These issues can be alleviated with the choice of an appropriate normalisation method.

Normalisation

- Dialysis efficiency is usually determined by URR
- Consider a relative change in VOC concentration.
- Normalise to half of the average of the two concentrations

$$C_{j,norm} = \frac{C_j}{C_j + C_k}$$

$$C_{k,norm} = \frac{C_k}{C_j + C_k}$$



- Data is bounded between [0 1]
  - > 0.5 = decrease in concentration over course of treatment
  - <0.5 = increase in concentration.

Biomarkers

- Select only masses with >0.5 non-zeros in either group
- Classify using selected biomarkers for sensitivity/specificity of that marker
- Biomarkers are displayed visually on an image plot using the difference between the pre- and post-dialysis groups.
- Relative change, *Δrel*

$$\Delta_{rel} = \frac{C_j}{C_j + C_k} - \frac{C_k}{C_j + C_k}$$

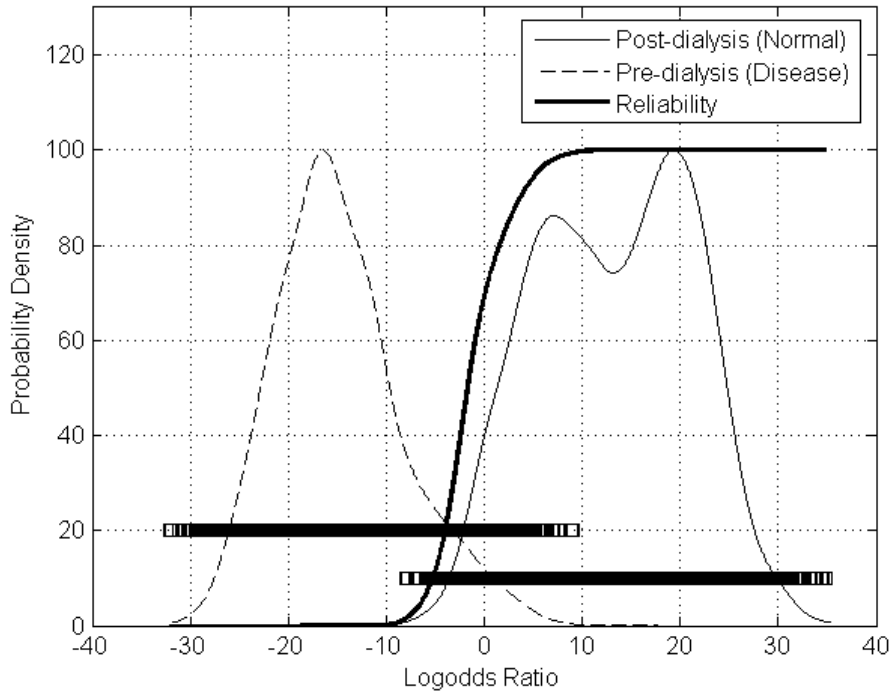
$$\Delta_{rel} = \frac{C_j - C_k}{C_j + C_k}$$

Biomarkers

- Density profiles are created from the means of successful bootstrap samples using *Δrel* and are combined into an image plot.
- The distribution of the mean is the same as the distribution of the raw data
  - Create image plot off means for easier marker ID
- Select best biomarkers: highest mean to standard deviation ratio.
  - Masses with narrow distribution centered far from increase/decrease interface
  - Crossing increase/decrease interface: bad biomarker

Dialysis proof-of-concept

- Raw data: sample size 39 before/after pairs
- Log-odds for minimum overlap
- Relative change for biomarker ID



ROC area ~ 0.995

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UCBIOENGINEERING

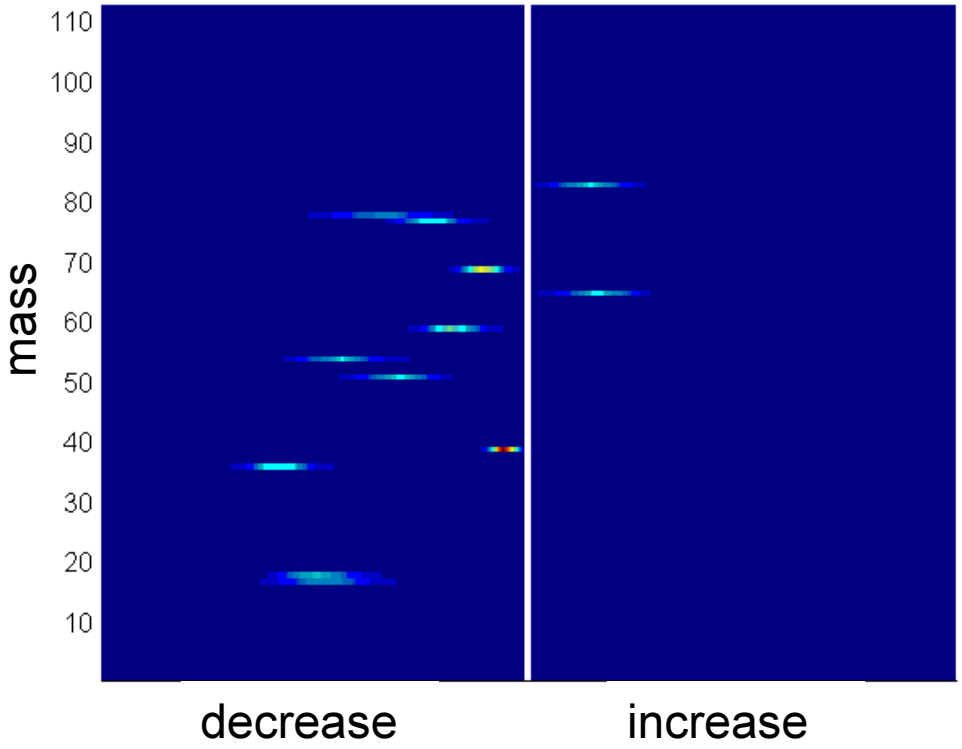
Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Biomarker Image plot

- Best biomarker: largest consistent relative change in concentration
- Masses centered far from the decrease/increase interface with the narrowest distributions



Classification Method		Biomarker Selection Method	
Mass	Score	Mass	Score
36	0.1478	36	1.888
18	0.2224	18	1.2933
54	0.3562	54	1.0517
77	0.4821	17	0.9732
51	0.4875	51	0.7781
59	0.5041	77	0.6339
17	0.5435	59	0.6309

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Summary

- **Two classification methods developed and validated**:
  - **Pure classification** = assumes patients are independent and requires many patients and repeats
    - Tedlar bag validation
    - Dialysis proof-of-concept, but small sample size
  - **Paired classification** = allows patients to be paired for classification rather than used independently, which effectively increases available data
    - Dialysis study presented gave better results this way

- **Methods interact with study design**
  - Number of patients and repeats per patient (i.e. quality and density of data you can get) will affect choice of method
  - Emphasises integrated study and classifier design

Acknowledgements

- **Dr. Dominic Lee and Dr. Geoff Chase at UoC**
- **Dr. Jenny Scotter, Dr. Randall Allardyce and Syft Technologies**
- **Dr. Zoltan Endre (dialysis study and data)**
- **Dr. Andrew Moot (tedlar bag study)**
- **TEC Top Achiever Doctoral Scholarship program**

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Questions or Comments?


UC
UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND


UCBIOENGINEERING

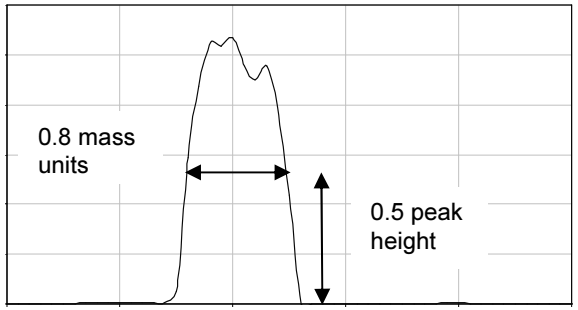

Canterbury
District Health Board
Te Poari Hauora ō Waitaha


Tertiary Education Commission
Te Amorangi Mātauranga Matua

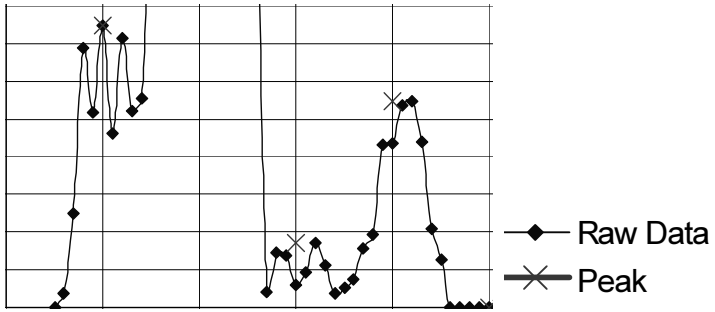

syft
Technologies

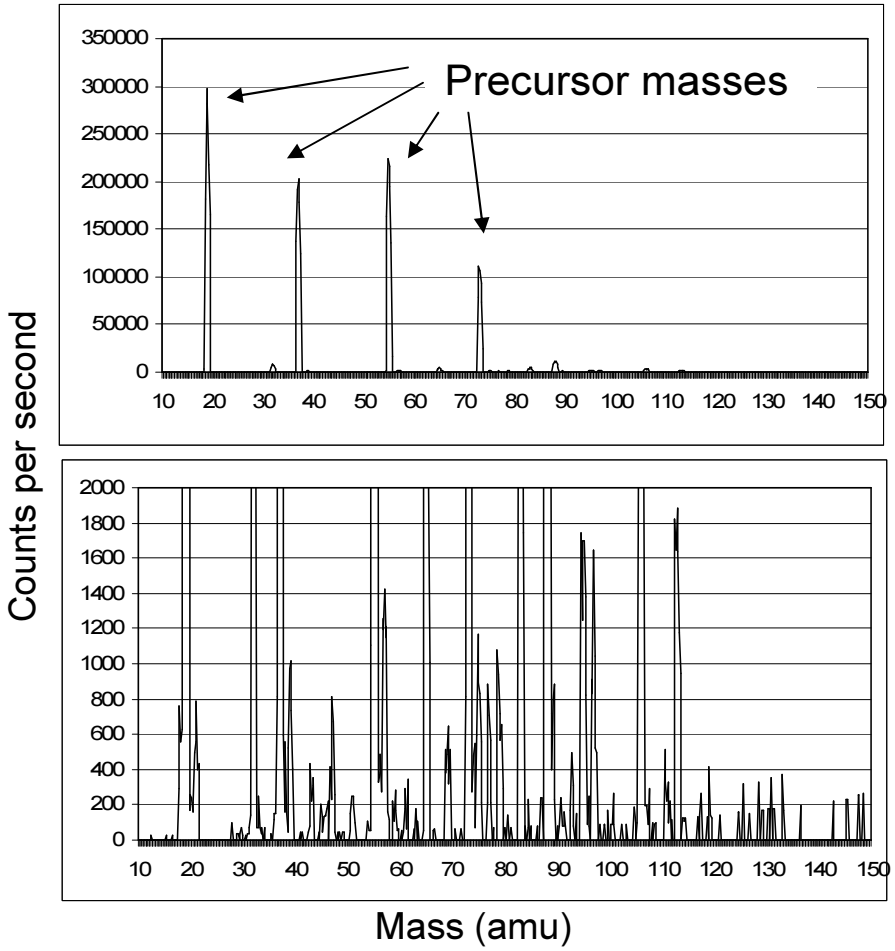
SIFT-MS

- SIM scans and Mass Scans
- Scans taken every 0.2 mass unit



- Noise
- Reduce to concentration matrix for whole mass units





UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies

Validation Study "Biomarkers"

Product ions	Mass	Explanation	Rank
H_3O^+ and its water clusters	19, 55, 73	The wet nitrogen group showed much higher concentrations at masses 55 (and 73), corresponding to the water clusters of H_3O^+.	3, 4, 5
Isotope of H_3O^+ and its water clusters	57	The wet nitrogen group showed much higher concentrations at mass 57, corresponding to the water clusters of H_3O^+, (and lower concentrations at mass 21, corresponding to the mass of the H_3O^+ isotope (D_3O^+) with no water cluster).	2
$C_4H_9NO.H^+$ (product of N,N-dimethyl acetamide	88	Due to the solubility of $C_4H_9NO.H^+$ in water, by venting the nitrogen through the water bottle, the concentration at mass 88 decreased dramatically, (and increases at mass 106 – its water cluster).	7
$C_6H_6O.H^+$ (product of phenol)	95	Due to the solubility of $C_6H_6O.H^+$ in water, by venting the nitrogen through the water bottle, the concentration at mass 95 decreased dramatically.	6
$N_2H^+.H_2O$	47	Water cluster of N_2H more prevalent when N_2 passes through water.	1

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

UC BIOENGINEERING

Canterbury District Health Board Te Poari Hauora ō Waitaha

Tertiary Education Commission Te Amorangi Mātauranga Matua

syft™ Technologies